March 27, 2020

MEMORANDUM FOR Dr. Ron S. Jarmin
Deputy Director and Chief Operating Officer

From: Dr. Victoria A. Velkoff
Associate Director for Demographic Programs

John M. Abowd

Dr. John M. Abowd
Associate Director for Research and Methodology
and Chief Scientist

Consulted with: Jason Devine, Christa Jones, Enrique Lamas, V. Thomas Mule, Matthew Spence, Deborah Stempowski

Subject: Estimating the Undocumented Population by State for Use in Apportionment

This memo is in response to your request to consider ways in which the Census Bureau could produce estimates of the undocumented foreign-born population that could be used to adjust the apportionment numbers for 2020. We have outlined two methods below. One method relies on an aggregate residual method similar to the one external researchers use to estimate the undocumented population. The second method relies on using administrative records and survey sources, including the American Community Survey (ACS) and coverage measurement surveys, matched to the 2020 Census person-level results to create models for predicting the undocumented population.

Historically, the Census Bureau has conveyed the enumerated population counts from the decennial census to the Secretary of Commerce for use by the President in Congressional apportionment, as outlined in the 1929 Reapportionment Act and the relevant sections of Title 13. The methods for estimating the undocumented population discussed below are based on various non-decennial data sources including administrative records and sample surveys. The results from all of the methods depend greatly on untestable assumptions as well as the choices of which data sets to incorporate. The most responsible conveyance of these results would include a range of estimates to reflect the uncertainty arising from untestable assumptions and the use of sample-based data.

It is our professional judgement that these methods produce estimates of the undocumented foreign-born population at the state level that could inform policy makers, but they could fail the prohibition of sample-based methods for the apportionment of the House of Representatives contained in Title 13, Section 195. Our goal is to produce a credible and accurate estimate of the undocumented foreign-born population that is objective and protects the confidentiality of the underlying micro-data. . If we are directed to produce an estimate of the undocumented population by state, we recommend using the aggregate method because we could begin work as soon as we had an instruction to do so. The second method requires using the 2020 Census Unedited File (CUF) which is schedule to be complete by November 30, 2020. The Census Bureau is scheduled to begin work on apportionment on December 7, 2020. We do not think that a week is enough time to implement the record linkage and statistical modeling required for a production use of this method. In addition, because of the inherent limitations of administrative records for resolving the undocumented status of respondents to the 2020 Census, a statistically valid implementation of the record linkage and modeling solution involves the use of the same untestable assumptions and supplemental data source choices as the aggregate method must use.

Note that Census Bureau has not researched either of these methods. In order to deliver by December 30, 2020, we would need to begin such a research program as soon as possible.

There is more detail below about the methods, the data sources, the assumptions, and the schedule.

Techniques to Estimate the Undocumented Population

The population counts that provide the basis for apportionment of the House of Representatives are currently based on an enumeration of all persons residing in each state. Estimating the undocumented foreign-born population included in this actual enumeration, in order to remove them from the apportionment resident population counts, would require a combination of historically-vetted demographic techniques and coverage measurement estimates with entirely new administrative data methods. These methods rely on difficult-to-verify assumptions and sample-based estimates that introduce substantial imprecision to enumerated population counts.

Two basic methods are available for estimating the number of undocumented immigrants included in the census enumeration of resident population. The aggregate method creates an estimate of the foreign born in the United States and then subtracts an estimated aggregate foreign-born population living legally in the United States to arrive at an estimated undocumented immigrant population. This method uses aggregates from administrative data, household surveys (e.g., the American Community Survey), and census coverage-measurement surveys to estimate the undocumented immigrant population counted in the census. The micro-data method uses record linkage between the census responses themselves and administrative data containing information on citizenship status and legal-resident alien status. The residual in this method is the actual enumerations that cannot be linked to any administrative source for documented citizenship status. This residual must be modeled into components that include citizens and documented resident aliens for whom the record linkage failed and a component that is the estimated undocumented count. Because there is very limited direct evidence on undocumented status, the models used to estimate these components also rely on demographic analysis, household and coverage-measurement surveys, and are subject to significant uncertainty.

1. Aggregate Residual Method

The aggregate residual method consists of creating an estimate of the foreign-born population in the United States (usually from the American Community Survey) and then estimating the number of foreign born who are in the United States legally. The estimate of the undocumented is the residual from subtracting the estimated legal resident foreign-born population from the estimated total foreign-born population.

First, the foreign-born population is estimated. This estimate would most likely come from the 2019 American Community Survey (ACS), since the ACS asks respondents about nativity and has a large sample size that can provide reliable state-level estimates. The 2020 ACS data are still being collected in December 2020 and would not be available to meet the December 31, 2020 timeline for disseminating apportionment counts. The ACS is a sample survey, and as such, this estimate of the foreign-born population is subject to sampling error. It is also subject to coverage, and incomplete data, and disclosure avoidance error.

The ACS estimates must be adjusted to change the reference date from July 1, 2019 (the mid-point for the 1-year 2019 ACS data) to April 1, 2020 (the reference date for the 2020 Census). To make this adjustment, we would need to estimate the number of migrants who arrive between these dates. We would also need to estimate the existing stock of foreign born persons who die, emigrate from the United States, or move between U.S. states, between July 1, 2019 and April 1, 2020. While we can create estimates of mortality from health records, there are limited data available to estimate emigration of the foreign born and the movement of the foreign born between states. These estimates will vary greatly depending upon the assumptions used, none of which can be easily verified empirically. For example, in estimating the movement of the foreign born between states, one might extrapolate from the most recent one-year average out-migration rate or from a multi-year average. In a recent study, the out-migration of foreign born in Alaska using a multi-year average was double that of the single-year average.

Next, several adjustments to the foreign-born population estimates are made. First, since the 1986 Immigration Reform and Control Act gave legal status to most undocumented immigrants who arrived prior to 1982, studies that use the aggregate residual method include only those foreign born who arrived after 1980. Using the ACS data permit an estimate of this component of the foreign-born population, since the survey asks foreign-born respondents about their year of entry. In addition to sampling error, there are known data quality issues with the ACS year of entry statistics, such as a high allocation rate for missing data and year heaping in the reported data, which make this adjustment subject to additional measurement error.

Finally, the estimates of the foreign born derived from the ACS should be adjusted to account for differential coverage rates. Prior work by the Census Bureau has shown that there is a difference in coverage between the ACS (a survey) and the decennial census (an enumeration) (Jensen, et al., 2015). We expect that the coverage rate will differ, because the decennial census has more media and advertising coverage. The decennial census is also available in more languages and has more in-person interviews (prior research has shown the foreign born are more likely to respond to an in-person interview than the native born). The aggregate residual method requires estimating these survey-to-census differential coverage rates of the foreign born and applying these differential rates to the state-by-state foreign-born population estimates.

Given the estimates of the foreign-born population in each state, discussed above, the second step of the residual method is to estimate the population of legal foreign-born residents. This estimate has traditionally come from the Department of Homeland Security’s Office of Immigration Statistics, which tabulates data on flows of new naturalizations, legal permanent residents, asylees, and refugees at the state level by year. These flows will require adjustment to account for deaths, emigration from the United States, and emigration to other states within the

United States to create an estimate of total legal foreign-born population by current state of residence. Estimates of the number of legal, temporary migrants must then be added to those state totals to get the total number of legal foreign-born residents for each state. As noted for the total foreign-born population, estimating each of these components of population change requires making assumptions that can have significant effects on that component's magnitude. For example, Appendix A shows three estimates of the annual emigration rates among a sizeable portion of the foreign born—male Mexican-born recent arrivals. These rates vary significantly, depending on the study's assumptions and estimates. These component effects compound into profound downstream changes on the estimated total size of the population.

The final step of the residual method is to subtract the estimated number of legal residents who are probable decennial census respondents from the estimated number of all foreign born who are probable decennial census respondents to get an estimate of the number of undocumented respondents to the 2020 Census. This number would then be subtracted from the total enumerated resident population counts.

Prior work on this method by the Census Bureau and other researchers have shown the challenges associated with this method. Woodrow (1991) uses the aggregate residual method to estimate the undocumented population for the 1990 Census. Minor differences in assumptions about sampling, coverage, or estimation lead to a wide range of estimates, such that the undocumented population in 1990 may have been as low as 1.5 million or as high as 4.5 million. Ahmed (1995) also uses the aggregate residual method and estimates the undocumented population to be 942,000 persons.

Several more recent studies have used the aggregate residual method to estimate the undocumented immigrant population. The Pew Research Center (Pew 2016), the Center for Migration Statistics (Warren 2016), and the Office of Immigration Statistics (OIS 2018) estimate the undocumented immigrant population as 10.7 million, 10.8 million, and 11.7 million, respectively (See Appendix C). Despite using very similar approaches, the final estimates differ by over 1 million people nationally.

The studies that have attempted to allocate the undocumented population among states illustrate the difficulties and shortcomings of this approach.

Fernandez and Robinson (1994) estimate the number of undocumented immigrants at the state level for the 1990 Census. To derive their estimate, they assume that the state distributions of undocumented persons "followed a pattern similar to that of foreign born non-citizen immigrants enumerated in the 1990 census by state." Even with this simplifying assumption, the authors provide a range of estimates that differ greatly depending on small differences in assumptions. For example, Texas may have had as few as 300,000 undocumented immigrants in 1990, or as many as 427,000. Nearly every state was assigned an estimated range that greatly exceeds the difference in populations between the state that receives the 435th representative in apportionment and the next runner-up.

Pew (2016) creates an estimate for the top six states and then comes up with a way to distribute the remainder to the balance of the states. Warren (2016) collapses eight states into a remainder, and OIS (2015) creates estimates for only the ten states that are estimated to have the highest number of undocumented immigrant residents. As shown in Appendix C, the estimates vary significantly between states – California is estimated to have 2.2, 2.5, or 2.9 million undocumented immigrant residents, depending on the study used.

Pew is careful to report that all estimate for the undocumented population are presented as rounded numbers to avoid the appearance of unwarranted precision. The rounding rules vary by size of the estimate and by data source. Pew's rounding rules are as follows:

Estimate greater than 10 million	Nearest 100,000
1 million to 10 million	Nearest 50,000
250,000 to 1 million	Nearest 25,000
100,000 to 250,000	Nearest 10,000

ACS-Based 5k to 100,000	Nearest 5,000
CPS-based 10k to 100,000	Nearest 5,000
ACS-based less than 5,000	Shown as less than 5,000
CPS-based less than 10,000	Shown as less than 10,000

In 2000, the last congressional seat was decided by less than 1,000 people which is well below the rounding rules stated above.

Pew also publishes a range for its estimate. For instance, Pew estimates that California, the state with the largest estimated undocumented population has 2.2 million undocumented residents in 2016 plus or minus 60,000. Pew estimates that Alabama's undocumented population was 55,000 in 2016, plus or minus 10,000. In other words, Pew estimates that Alabama's undocumented population could be as low as 45,000 or as high as 65,000. For Alaska, Pew estimates the undocumented population is 5,000 with a range of 5,000. Alaska could have as many as 10,000 undocumented or none.

The OIS (2019) report illustrates some of the assumptions that must be made in doing this analysis. For example, OIS assumes that the current state of residence for a foreign-born individual is the same as it was when that individual applied for Legal Permanent Residence or Naturalization. This "ignores subsequent internal migration and affects the state-level estimates" (OIS 2019). OIS uses "a three-year moving average … for year of entry to reduce heaping effects" (OIS 2019). Additionally, OIS measures derived citizenship – which occurs when a non-citizen child under 18 becomes a citizen upon the naturalization of a parent – by looking at applications for naturalization certificates, even though this is not required by law, and many derived citizens opt to apply for U.S. passports as proof of citizenship instead. As the report notes, "derivative citizens may not request a certificate due to the high filing fee (currently $1,170, compared to $65 for a passport card, which also proves citizenship)." This method will lead to an undercount of derived citizens (OIS 2019).

A final challenge in this methodology is that the aggregate residual not only includes undocumented immigrants, but also includes people with legal status who are not yet included in the official estimates of legal migrants and refugees and people in "quasi-legal" status who are awaiting action on their legal migration requests. This leads to an upward bias in the estimates of undocumented immigrants, which may also vary by state in magnitude.

2. Record Linkage and Modeling Residual Method

An alternative method of estimating the undocumented immigrant population is to link decennial census responses to administrative data that indicate documented citizenship status. This method uses the actual 2020 Census responses to determine the resident population of each state. It uses the 2020 Census production record linkage system, as implemented for the other administrative record uses in this decennial. However, it also has a residual—the 2020 Census person records for which the administrative records cannot resolve citizenship status and the person records that cannot be matched to any administrative data. For the persons whose citizenship status cannot be determined by direct record linkage, we would use a statistical model to impute citizenship status, including documented v. undocumented status for non-citizenship. The principal caveat to this use of statistical imputation is that it cannot rely exclusively on information provided by the 2020 Census, including administrative data linked to the person records. The ensemble of data available in the combined 2020 Census person record and linked administrative records on documented citizenship status are essentially uninformative about undocumented status.

In addition to the technical issues, there are also administrative challenges. The Census Bureau currently has Memoranda of Understanding (MOUs) with the agencies that supply the administrative data. The MOUs include clauses delimiting the uses to which we may put those data. The specific permitted uses are detailed in each MOU. Those permitted uses are currently limited to determining citizenship status only and do not include resolving the documented status of non-citizens. The Census Bureau would need to seek modification of these MOUs in order to use the records to disaggregate non-citizen status into documented and undocumented categories, as required in the method presented in this section.

2.a. Ingestion of administrative data

The first step in the record linkage method is to ingest administrative data that can be used to resolve the citizenship status of persons living in the United States. These data determine the information available to augment the response data from the 2020 Census with a determination of citizenship. The list is of available sources is shown in Table 1, ordered according to the priority that the Census Bureau gave to that source's data for the purpose of resoling citizen v. non-citizen status.

These data were ingested, consistent with Executive Order 13880, although many of the sources, particularly the first two, were already being used as part of the 2020 Census administrative record program. The ongoing research using these data supports producing estimates of the citizen voting-age population by race and ethnicity at the census block level, as documented on the CVAP page of census.gov. To that end, the research team has used methods that produce a binary variable "citizen/non-citizen" for those persons believed to be adults as of April 1, 2020. These administrative data directly support the CVAP use case, as the third column of Table 1 clarifies.

The primary statistical tool the research team is using accumulates the evidence from each source into a best-estimate probability of citizenship for each person in the administrative record universe. When two or more administrative sources strongly agree that a person is a citizen, the estimated probability is very close to unity. In this case, a business rule is developed that classifies such persons automatically as "citizen," without passing through the statistical model. This happens, for example, when the person is recorded as U.S.-born and citizen in the Numident, does not have an ITIN, and does not appear in the lawful permanent resident/naturalized citizen data. It also happens when a person appears in the Numident as foreign-born, non-citizen, but also appears in the naturalized citizen data. Rule-based non-citizen status (best citizenship probability of zero) occurs when a person, for example, appears in the Numident as foreign-born, non-citizen and appears in the legal permanent resident data.

When the administrative data provide conflicting evidence of the citizenship status of the individual, the statistical model assigns the best estimate of the probability of citizenship, a number between zero and unity, which can then be used to statistically impute citizen or non-citizen. Regardless of the method used to do the final imputation, the important feature of this process is that the administrative data themselves provide the information for the statistical model and the business-rule-based classification into citizen and non-citizen categories. This is because the administrative data are informative about these two categories.

The administrative data and the information on the 2020 Census questionnaire itself are not directly informative for subcategories of non-citizens. Documented non-citizens are those who appear in the administrative sources as legal permanent residents and visa holders. Undocumented foreign-born persons generally do not appear in the administrative records, with the exception of those who overstay their visas, which demographic estimates suggest is an increasing percentage of the administrative-record universe but still thought to be less than 50 percent of the undocumented foreign-born population (Warren, 2020). The most important consequence of the invisibility of the undocumented foreign-born population in the administrative records is that statistical models based on those data will not have very much, if any, useful information for classifying non-citizens into “documented” and “undocumented” categories. Such models—whether they use the traditional hot-deck imputation or the modern latent classification imputation—must be supplied with external information identical to the information used in the aggregate residual method. Those statistical models then reliably translate this external information into probabilities that a non-citizen is documented or undocumented. The imputation is therefore based primarily on the external data, and not primarily on either the administrative records or the 2020 Census information.

Table 1: Administrative records data sources available for status resolution and modeling, listed in priority order from the EO 13880 tracking

Social Security Administration (SSA)	Numident	Data on Country of Birth and Citizenship resolves U.S.-born and citizens. Does not completely resolve foreign born with missing citizenship or non-citizen foreign born.
Internal Revenue Service (IRS)	1040 and 1099 forms filed with Individual Taxpayer Identification Numbers (ITIN)	ITIN are only issued to non-citizens. This partially resolves legal status.
Department of Homeland Security (DHS)	a. Lawful Permanent Resident File and Naturalization Data (CIS)	Data for Legal Permanent Residents and Naturalizations resolves legal/documented status for those included.
	b. VISA Data (ICE)	Data show visa status for temporary migrants. Does not completely resolve documented status.
	c. Arrival-Departure Information System Data (CBP)	Data may be useful to track emigrations of non-citizens and partially resolve documented status.
Department of State (State)	Passport Services*	Data on passport holders resolves legal status for those included, in particular derived citizenship for children.
SSA	Master Beneficiary Records (MBR), Supplemental Security Records (SSR), Payment History Update System (PHUS)	Eligibility rules for these programs partially resolve legal status for those included.
Department of Health and Human Services (HHS)	Center for Medicare and Medicaid Services (CMS) Medicare, Medicaid, Children's Health Insurance Program (CHIP)	Eligibility rules for these programs partially resolve legal status for those included
Department of Justice	a. Bureau of Prisons	Contain citizenship status of prisoners and detainees.
	b. U.S. Marshals Service	

Department of Housing and Urban Development (HUD)	Federal Housing Administration, Public and Indian Housing Information Center, Tenant and Rental Assistance Certification System, Computerized Homes Underwriting Management System, Low Income Housing Tax Credits	Data from these systems contains information on payments and services delivered. Eligibility for these services and payments partially resolves citizenship status.
HHS	Indian Health Services*	Eligibility rules for these programs resolve citizenship status.
State	Worldwide Refugee Admissions Processing System	Data on asylees and refugees resolves legal status for those included.
Bureau of Justice Statistics (BJS)	National Corrections Reporting Programs*	Direct information resolving citizenship status.
Data from individual states	Temporary Assistance for Needy Familes (TANF), Supplemental Nutrition Assistance Programs (SNAP), Supplemental Nurition Program for Women, Infants and Children (WIC), Drivers Licenses**	Direct information and eligibility rules partially resolve legal status.
*These data had not been ingested by the Census Bureau as of March 8, 2020. **These data are not available for every state.		

2.b. Linkage between administrative and 2020 Census data

The second step in the record linkage procedure is to determine the individuals in the administrative data universe who successfully link to a 2020 Census person record. This record linkage step is materially different for the administrative universe as compared to the 2020 Census person records. The vast majority of the administrative records ingested by the Census Bureau, including all of the Numident and IRS data, contain a unique administrative identifier, usually a Social Security Number (SSN). In this case, the Bureau's production record linkage system—the Person Identification Validation System (PVS) —directly assigns the internal identifier—the Protected Identification Key (PIK)—without using any statistical record linkage methods. In this case, the false match and false non-match rates associated with PVS are both zero. The record linkage is exact.

When the PVS processes a 2020 Census person record from the Census Unedited File (CUF), there is no SSN available. PVS uses probabilistic record linkage to compare name, sex, address, and birth date (PII) to its master reference file. A successful match associates the PIK with the highest agreement score between the CUF record and the master reference file as long as the agreement score exceeds a statistically set cutoff. Internal Census Bureau research (Layne, Wagner and Rothhaas, 2014) shows that when PVS processes high quality PII, the false match rate is less than 0.005%, but when PVS processes low quality PII, the false match rate rises to more than 10%. External reviews of PVS (NORC, 2013) confirm this result. These results imply that when the citizenship status is linked from the administrative universe to the 2020 Census person records, the results are most reliable when the Census response contains high quality PII.

High quality PII on the Census come from self-responses and non-response follow-ups supplied by a household member. NRFU responses from proxies are much lower quality. The 2010 Census Coverage Measurement studies (Mule, 2012), which used a methodology similar to PVS, reported results that demonstrate the role of high quality PII. When the PII came from self-response, 96.9% of 2010 Census responses were correct—meaning that the PII on the Census questionnaire and the PII on coverage measurement survey successfully matched. When the PII came from a proxy response, only 70.1% of the Census questionnaires were correct—meaning that the PII failed to match the coverage measurement in 29.9% of these cases.

When the PVS processes a CUF record, the quality of the PII determine both the overall match rate and the expected false match rate in the resulting data. The table in Appendix B shows that when PVS processed the 2010 CUF, the national match rate was 88.6%, but the state-level match rates varied from a high of 94.1% in North Dakota to a low of 83.4% in Nevada. The records that link have expected false match rates that very from less than 0.005% (when the PII is high quality) to more than 10% (when the PII is low quality). If the 2010 Census results generalize to the 2020 Census, we can expect about 95% of the matches (self-responses plus NRFU responses by a household member) to produce high quality PII, with the balance (NRFU proxy responses and unresolved) producing low quality PII. The extent and quality of this record linkage directly determines how useful the administrative universe data will be in providing direct evidence of citizenship and how much statistical imputation will be needed for the unlinked 2020 Census persons.

2.c. Estimating the undocumented foreign-born population by state

A very large fraction of the persons in the CUF who link to the administrative universe will have their citizenship status (citizen, documented non-citizen, undocumented non-citizen) resolved from the ensemble of the administrative sources in Table 1. Those assigned by weight-of-the-evidence business rules will have the value of one in one of these three categories and zero for the other two. The remainder of the linked cases will have a probability between zero and one (summing to one) for each category that can be used for statistical imputation.

The CUF persons with resolved citizenship from the administrative universe are used to estimate statistical models for imputation of the citizenship status of the unlinked CUF persons. There are at least two feasible methods for modeling this imputation, but they both embody similar statistical assumptions. The methods can be based on the assumption of "ignorable missing data," which means that, given the characteristics observed in the 2020 Census (location, sex, age, race, ethnicity, composition of household), the probability that an unlinked CUF person is citizen, documented non-citizen, or undocumented non-citizen is the same as the probability of a linked person with the same characteristics. All missing data models used in the 2020 Census production system to produce the Census Edited File embody this assumption. In the context of imputing missing citizenship status, the ignorable missing data assumption implies that a person in the unlinked CUF records is just as likely to be an undocumented immigrant as a similar person in the linked records.

The demographic analysis discussed in the aggregate residual method strongly suggests that persons in the unlinked CUF records are less likely to be citizens and, among non-citizens, more likely to be undocumented. That is, the demographic analysis documents why the ignorable missing data assumption is untenable for imputing citizenship status among the unlinked CUF persons. Furthermore, the unlinked persons are less likely to be citizens because the reference list in the PVS system is based on documents that citizens are more likely to possess. The unlinked persons are more likely to be undocumented immigrants because it is precisely the documents that confirm legal resident alien status that were used to resolve citizenship status among the linked persons. If a person has no documents, that person's probability of being in the administrative record universe is much lower, approaching zero if the person has had no interaction with any part of the federal or state agencies covered by the records in Table 1.

To summarize, the "residual" in the record linkage method consists of all CUF persons whose citizenship status was not resolved by the administrative records. We call these persons "unlinked." Statistical modeling must be used to impute their citizenship status. These models come in two types: those that make the "ignorable missing data" assumption, and those that do not. Models that do not assume that the missing data are ignorable, must incorporate extra-Census information in order to adjust the probabilities of citizen, documented non-citizen and undocumented non-citizen away from those implied by the ignorable missing data assumption. These adjustments are based on exactly the same demographic analysis that was used in the aggregate residual method. Specifically, estimates of the proportion of the foreign-born population who are undocumented, given location, sex, age, race, ethnicity, and household composition.

2.c.1. Hot-deck imputation of missing citizenship status

The most common statistical imputation method used in the 2020 Census is a hot deck system. To implement a hot deck, the CUF records with resolved citizenship status are sorted into bins that aggregate location, sex, age, race, ethnicity, and household composition according to rules that specify the minimum size of each bin. This is called the hot deck matrix. Then, each CUF record with unresolved citizenship status is matched with the correct bin in the hot-deck matrix for that record. One of the resolved CUF records in that bin is randomly selected. Citizenship status from the randomly selected record is imputed to the record with the missing status. Hot-deck imputation implemented by this procedure insures that the probabilities used to impute citizenship match the probabilities in the linked CUF records.

To implement hot-deck imputation with non-ignorable missing data, each cell in the hot-deck matrix must be assigned a set of adjustment factors. These factors are based on demographic estimates of how much less likely a person is to be a citizen or documented non-citizen, given their unlinked status and the other characteristics. These probabilities would be estimated from the American Community Survey and Coverage Measurement Survey as described in the aggregate residual method. Unless this adjustment is made, the hot-deck imputation would deliver a downward biased estimate of the undocumented foreign-born population that would vary by state and other important demographic characteristics.

2.c.2. Statistical models for imputing citizenship status

The statistical model used to resolve citizenship status among the linked CUF records can be extended to impute citizenship status for the unlinked persons. To implement model-based imputation in this framework, the Census Bureau would have to specify the prior information available to adjust the probabilities. If no prior information is used, the statistical model implements the ignorable missing data assumption, and its results would be very similar to the hot-deck method. Specifically, it would be expected to under-estimate the undocumented foreign-born population.

These statistical models can incorporate data-driven supplemental information from sample surveys or other data sources. Census Bureau could estimate the probability of being a citizen, a documented foreign-born person, or an undocumented immigrant based on this information. For example, Bachmeier et al. (2014) and van Hook et al. (2015) suggest using responses from the Survey of Income and Program Participation (SIPP) to develop a demographic model that imputes legal status for ACS responses. Although this method has only limited testing in another context, , it appears to be feasible. The ACS and SIPP questionnaires permit a much richer set of predictors, such as citizenship status and place of birth, which can be used to model legal status; however, only variables that are collected in 2020 Census questionnaire, which is much more limited, can be used in these model. Additionally, SIPP responses may be less generalizable to the decennial census than ACS responses, because decennial responses include many more hard-to-count individuals.

2.d. Limitations of the imputation methods

Formally, all other statistical imputation models used in the 2020 Census rely on the assumption of ignorable missing data. This assumption implies that the complete data records (in this case, those that include documented citizenship status) can be used to predict the incomplete data records (in this case, those missing documented citizenship status) without relying on extra-census data. For example, when implementing count imputation for housing unit addresses not resolved in nonresponse followup, the statistical model randomly selects a household that was enumerated to estimate the number of persons living in the unresolved address. A similar model for the persons whose documented citizenship status has not been resolved by the linked administrative data would select the status from a person whose status was resolved. This means that the only undocumented foreign-born persons this method could impute would be based on observed rate of persons overstaying their visas among 2020 respondents whose data were linked. We expect this rate to be less than half of the undocumented population. By contrast, when the statistical models are augmented using data similar to the demographic analysis described in Section 1, they produce reasonable, but again highly variable, estimates of the undocumented foreign-born population. Continuing the analogy to count imputation, we do not augment the count imputation model for the 2020 Census with any non-decennial data. If we used the post-enumeration survey estimates to adjust the count imputation for net under-coverage, instead of using the number of persons in a randomly selected respondent household, we would multiply that number by the coverage adjustment factor implied by the post-enumeration survey.

A second major challenge in implementing any model is that item-level edits and characteristic imputation within the 2020 Census data will not be available before the apportionment counts are delivered. Therefore, any model will have to account for potentially inconsistent data and patterns of non-ignorable missing data.

There is also a scheduling challenge with any imputation procedure. The record linkage and modeling approach uses confidential data protected by U.S. Code Title 13, Section 9. The resulting state-level estimates of the undocumented immigrant population must be processed by the differentially private 2020 Census Disclosure Avoidance System (DAS) using the share of the privacy-loss budget assigned to this tabulation. This disclosure avoidance methodology injects random noise into the tabulations to protect respondents' information and identity. The amount of noise injected will be determined by the global privacy-loss budget set for all 2020 Census data products, and therefore protecting the undocumented immigrant population counts will negatively affect the accuracy of other 2020 Census data products, such as the Public Law 94-171 redistricting files.

3. Timeline and resources

Apportionment counts--one number for each state, Washington D.C., and Puerto Rico--are due to the President no later than December 31, 2020. In order to support the activities and processes necessary for the tabulation of those counts, the data collected through self-response and nonresponse follow-up (in-field work) go through a variety of processing steps. These post-data collection processing activities begin in earnest once all in-field work is completed (expected by July 31, 2020 according to the schedule). Although completion dates for in-field work are planned carefully, unforeseen events, such as natural disasters and epidemics, can extend the data collection period. During both the 2010 Census and 2000 Census, data collection in-field work did not finish until late August to accommodate late operational needs.

At the completion of in-field work and geographic processing, the creation of the Decennial Response File #1 (DRF1) begins.   DRF1 must be delivered to downstream processes as input to tabulation no later than October 14. In preparation for that delivery, the post-data collection integration process involves a variety of activities, including the following: (1) determining the final disposition of suspected fraud cases, (2) matching addresses and removing duplicate addresses (unduplication) to incorporate the final census updates into the master address file, (3) completing the coding of residence locations, and (4) unduplicating persons within a census return to ensure the accuracy of household rosters. These technical processes and programs are overseen by decennial census experts, such as fraud detection analysts, geographers, demographers, and mathematical statisticians. At the end of these processes, 52 files are created, one for each state, the District of Columbia, and Puerto Rico. Expert demographers review the 52 files on a flow basis over a short three-week period. If errors are found during the review, the computer programs must be re-run in part or in whole. Shortening the review cycle or processing time would increase the risk that the Census Bureau will not successfully produce high quality final population counts, and could produce subsequent delays in processing the data and producing mandatory data products.

Once the DRF1 is complete, work begins on the Decennial Response File #2 (DRF2). Among other activities, the Primary Selection Algorithm (PSA) is applied in the creation of DRF2. The PSA is a complex program used to determine who should be counted at each residence from which two or more responses have been collected. Statistical experts within the decennial census programs have carefully studied many potential scenarios and reviewed data from past censuses and census tests to formulate the PSA. Details of the PSA are extremely sensitive and thus are known to only a small group of experts inside the Census Bureau. The PSA must be run on a file that includes all responses across the United States, and takes approximately five days to run. DRF2 is scheduled for completion and delivery to downstream processes on November 4, 2020.

Similar to DRF1, 52 files are created, one for each state, the District of Columbia, and Puerto Rico. Expert mathematical statisticians and demographers perform their review on a flow basis across the 52 files over a short period. As with the review of DRF1, shortening the process or review time for DRF2 would increase the risk of errors in the data or delays in the schedule. Potential implications include duplicating people in the census and additional runs of the processes if errors in the data are found downstream.

After completion, DRF2 receives additional processing to create the Census Unedited File (CUF), which serves as the basis for apportionment, among other purposes. This additional processing to create the CUF ensures that the universe of Census data is complete, and that the state portion of the geospatial identifier can no longer be edited. The CUF will be completed by November 30. Subject matter experts must conduct a thorough review of the counts and data on the CUF to check for accuracy and completeness, then tally state populations and perform the apportionment of Congressional seats by December 31, 2020. In this same time frame, staff plan for the necessary reruns of programs to address errors before the December 31, 2020 deadline.

There is a long series of necessary, but complicated, steps that must be conducted to complete this process. The steps include running very detailed and complex computer programs, and extensive review of the results by subject matter experts. Any problems in the sequence that lead to a delay can affect the entire remaining schedule. However, the deadline for delivering the apportionment counts cannot be moved beyond the statutory due date, December 31, 2020. Any significant changes to the current processing and its schedule place that delivery at serious risk.

References

Ahmed, Bashirudden. unpublished. “Estimates of undocumented immigration counted in the 1990 Census.” (Draft: March) (U.S. Census Bureau, Population Division). Cited in Costanzo, Joseph, Cynthia Davis, Caribert Irazi, Daniel Goodkind, and Roberto Ramirez (2001) “Evaluating Components of International Migration: The Residual Foreign Born,” Population Division Working Paper No. 61.

Bachmeier, James D., Jennifer Van Hook, and Frank D. Bean. 2014. “Can We Measure Immigrants’ Legal Status? Lessons from Two U.S. Surveys.” International Migration Review 48(2), 538-566.

Department of Homeland Security, March 5, 2019, “Potential Improvements to DHS Illegal Alien Population Estimates: Collection and Use of Data,” Fiscal Year 2018 Report to Congress.

Fernandez, Edward and J. Gregory Robinson (1994) “Illustrative Ranges of the Distribution of Undocumented Immigrants by State,” Population Division Working Paper No. 8.

Jensen, Eric B., Renuka Bhaskar, and Melissa Scopilliti, June 2015, “Demographic Analysis 2010: Estimates of the Coverage of the Foreign-Born Population in the American Community Survey,” Population Division Working Paper No. 103.

Layne, Mary, Deborah Wagner and Cynthia Rothhaas, 2014 “Estimating Record Linkage False Match Rate for the Person Identification Validation System,” Center for Administrative Records Research and Applications Working Paper No. 2014-02.

Mule, V. Thomas (2012) “Census Coverage Measurement Estimation Report: Summary of Estimates of Coverage for Persons in the United States,” Decennial Statistical Studies Division 2010 Census Coverage Measurement Memorandum Series #2010-G-01.

NORC (2013) “PVS Research: Task 4, Further PVS Research Final Research Report,” Technical report delivered to the U.S. Census Bureau March 15, 2013.

Office of Immigration Statistics, December 2018, “Population Estimates: Illegal Alien Population Residing in the United States: January 2015,” Department of Homeland Security, Office of Strategy, Policy and Plans.

Office of Immigration Statistics, May 2019, “Population Estimates: Lawful Permanent Residents in the United States: January 2015,” Department of Homeland Security, Office of Strategy, Policy and Plans.

Pew Research Center, November 27, 2018, “Unauthorized Immigrant Total Dips to Lowest Level in a Decade.”

Rastogi, Sonya and Amy O’Hara (2012) 2010 Census Match Study Final Report. 2010 Census Planning Memoranda Series No. 247, Issued November 19, 2012

Van Hook, Jennifer, and James D. Bachmeier. 2013. "How Well Does the American Community Survey Count Naturalized Citizens?" Demographic Research 29(1), 1-32.

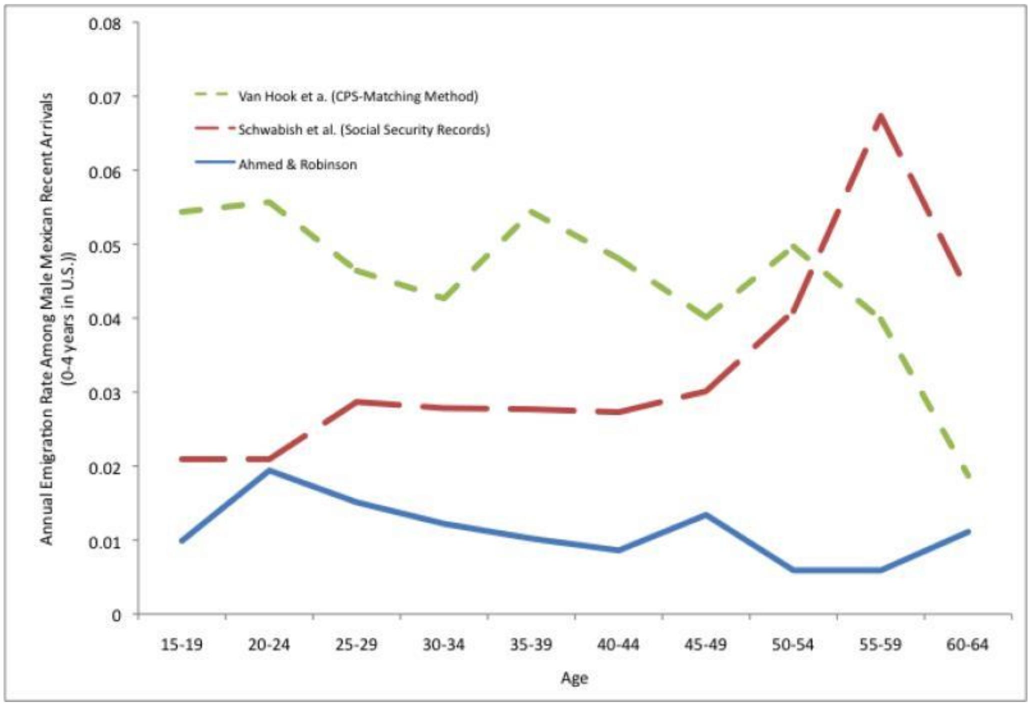
Van Hook, Jennifer, James D. Bachmeier, Donna Coffman, and Ofer Harel. 2015. "Can We Spin Straw Into Gold? An Evaluation of Immigrant Legal Status Imputation Approaches." Demography 52(1): 329-354.

Warren, Robert, 2018, 'The US Undocumented Population Fell Sharply During the Obama Administration," Center for Migration Studies.

Warren, Robert, 2020, "Reverse Migration to Mexico Led to US Undocumented Population Decline: 2010 to 2018," Journal on Migration and Human Security.

Woodrow, Karen (1991) "Preliminary Estimates of Undocumented Residents in 1990," Demographic Analysis (DA) Evaluation Project D2.

Appendix A: Three Estimates of the Annual Emigration Rate Among Male Mexican-Born Recent Arrivals to the U.S.



Source: van Hook and Bachmeier (2013).

Appendix B: 2010 Census and Administrative Records Match Ratios By State

State	Match Rate
North Dakota	94.1
Vermont	94.1
Iowa	93.6
Wisconsin	93.4
Maine	93.3
Minnesota	93.3
New Hampshire	93.3
Ohio	92.4
Nebraska	92.1
Pennsylvania	92.1
Kansas	92.0
Indiana	91.9
Missouri	91.7
South Dakota	91.6
Massachusetts	91.5
Connecticut	91.0
Kentucky	90.7
Rhode Island	90.6
Michigan	90.1
Virginia	90.1
Montana	90.0
Arkansas	89.9
Illinois	89.9
Tennessee	89.9
Utah	89.8
Washington	89.7
Mississippi	89.6
South Carolina	89.6
Alaska	89.5
West Virginia	89.4
Oregon	89.3
Wyoming	89.3
Maryland	89.2
Idaho	89.1
New Jersey	89.1
Oklahoma	89.1
Delaware	88.7
National	*88.6*
Alabama	88.5
Louisiana	88.2
North Carolina	87.9
Florida	87.7
Colorado	87.3

Hawaii	86.9
New York	86.8
Georgia	86.0
Texas	85.9
District of Columbia	85.0
New Mexico	85.0
California	84.8
Arizona	84.0
Nevada	83.4

Source: Rastogi and O’Hara (2012) 2010 Census Match Study, Table 9.

Appendix C: Various State Estimates of the Undocumented Immigrant Population

State of Residence	OIS (2015)	Pew (2016)	Warren (2016)
Alabama	N[1]	55,000	56,000
Alaska	N[1]	5,000	N[2]
Arizona	380,000	275,000	263,000
Arkansas	N[1]	55,000	50,000
California	2,880,000	2,200,000	2,548,000
Colorado	N[1]	190,000	184,000
Connecticut	N[1]	120,000	108,000
Delaware	N[1]	30,000	24,000
District of Columbia	N[1]	25,000	19,000
Florida	810,000	775,000	733,000
Georgia	390,000	400,000	353,000
Hawaii	N[1]	45,000	44,000
Idaho	N[1]	35,000	32,000
Illinois	450,000	400,000	476,000
Indiana	N[1]	100,000	96,000
Iowa	N[1]	50,000	48,000
Kansas	N[1]	75,000	77,000
Kentucky	N[1]	35,000	36,000
Louisiana	N[1]	70,000	62,000
Maine	N[1]	5,000	N[2]
Maryland	N[1]	275,000	235,000
Massachusetts	N[1]	250,000	159,000
Michigan	N[1]	100,000	101,000
Minnesota	N[1]	95,000	89,000
Mississippi	N[1]	20,000	18,000
Missouri	N[1]	60,000	54,000
Montana	N[1]	5,000	N[2]
Nebraska	N[1]	60,000	49,000
Nevada	N[1]	210,000	176,000
New Hampshire	N[1]	10,000	12,000
New Jersey	440,000	475,000	452,000
New Mexico	N[1]	60,000	56,000
New York	590,000	725,000	802,000
North Carolina	390,000	325,000	280,000
North Dakota	N[1]	5,000	N[2]
Ohio	N[1]	90,000	76,000
Oklahoma	N[1]	85,000	84,000
Oregon	N[1]	110,000	102,000
Pennsylvania	N[1]	170,000	159,000
Rhode Island	N[1]	30,000	24,000
South Carolina	N[1]	85,000	87,000
South Dakota	N[1]	5,000	N[2]
Tennessee	N[1]	130,000	118,000
Texas	1,940,000	1,600,000	1,758,000

Utah	N[1]	95,000	86,000
Vermont	N[1]	5,000	N[2]
Virginia	310,000	275,000	250,000
Washington	N[1]	240,000	242,000
West Virginia	N[1]	5,000	N[2]
Wisconsin	N[1]	75,000	75,000
Wyoming	N[1]	5,000	56,000
National Total	11,970,000	10,700,000	10,790,000

Source: Office of Immigration Statistics, 2015; Pew Research Center, 2016; Warren, 2016.

[1] OIS (2015) does not provide estimates for the states judged to be outside the top ten in terms of undocumented immigrant population. N represents these states.

[2] Warren (2016) does not provide estimates for these states.